Question **1**
Not yet answered
Marked out of 1.00

When constructing a *word embedding*, what is **true** regarding *negative samples*?

Select one:

- a. They are oversampled if less frequent
- b. They are words that do not appear as context words
- c. They are selected among words which are not stop words
- d. Their frequency is decreased down to its logarithm

Question **2**
Not yet answered
Marked out of 1.00

A page that points to *all other pages* but is not pointed by *any other page* would have...

Select one:

- a. Zero hub
- b. Nonzero authority
- c. Nonzero pagerank
- d. None of the above

Question **3**
Not yet answered
Marked out of 1.00

Considering the transaction below, which one is **false**?

Transaction ID	Items Bought
1	Tea
2	Tea, Yoghurt
3	Tea, Yoghurt, Kebap
4	Kebap
5	Tea, Kebap

Select one:

- a. {Yoghurt, Kebap} has 20% support
- b. {Yoghurt} has the lowest support among all itemsets
- c. {Yoghurt} -> {Kebab} has 50% confidence
- d. {Tea} has the highest support

Question **4**
Not yet answered
Marked out of 1.00

In *Ranked Retrieval*, the result at position k is *non-relevant* and at k+1 is *relevant*. Which of the following is **always true**?
Hint: P@k and R@k are the *precision* and *recall* of the result set consisting of the k top ranked documents.

Select one:

- a. P@k-1>P@k+1
- b. R@k-1<R@k+1
- c. R@k-1=R@k+1
- d. P@k-1=P@k+1

Question **5**

Not yet answered

Marked out of 1.00

What is **true** regarding *Fagin's algorithm*?

Select one:

- a. It provably returns the k documents with the largest aggregate scores
- b. It never reads more than $(kn)^{½}$ entries from a posting list
- c. It performs a complete scan over the posting files
- d. Posting files need to be indexed by TF-IDF weights

Question **6**

Not yet answered

Marked out of 1.00

Suppose that in a given *FP Tree*, an item in a leaf node N exists in every path. Which of the following is **true**?

Select one:

- a. The item N exists in every candidate set
- b. {N}'s minimum possible support is equal to the number of paths
- c. N co-occurs with its prefixes in every transaction
- d. For every node P that is a parent of N in the FP tree, confidence(P->N) = 1

Question **7**

Not yet answered

Marked out of 1.00

Which of the following is **false** regarding *K-means* and *DBSCAN*?

Select one:

- a. K-means does not handle outliers, while DBSCAN does
- b. K-means does many iterations, while DBSCAN does not
- c. K-means takes the number of clusters as parameter, while DBSCAN does not take any parameter
- d. Both are unsupervised

Question **8**

Not yet answered

Marked out of 1.00

Suppose that q is *density reachable* from p. The chain of points that ensure this relationship are {t,u,g,r}. Which of the following is **always true**?

Select one:

- a. p is a border point
- b. p is density reachable from q
- c. q and p are density-connected
- d. q is a core point

Question **9**

Not yet answered

Marked out of 1.00

Which of the following is **true** regarding *inverted files*?

Select one:

- a. The space requirement for the postings file is $O(n^{\beta})$, where β is generally between 0.4 and 0.6
- b. Inverted files prioritize efficiency on insertion over efficiency on search
- c. Storing differences among word addresses reduces the size of the postings file
- d. Compression by means of coding frequent values reduces the size of the index file

Question **10**

Not yet answered

Marked out of 1.00

Which attribute gives the **best** split?

A1	*P*	*N*
a	4	4
b	4	4

A2	*P*	*N*
x	5	1
y	3	3

A3	*P*	*N*
t	6	1
j	2	3

Select one:

- a. All the same
- b. A3
- c. A1
- d. A2

Question **11**

Not yet answered

Marked out of 1.00

Which of the following statements on *Latent Semantic Indexing* (LSI) and *Word Embeddings* (WE) is **false**?

Select one:

- a. LSI does not depend on the order of words in the document, whereas WE does
- b. LSI is deterministic (given the dimension), whereas WE is not
- c. The dimensions of LSI can be interpreted as concepts, whereas those of WE cannot
- d. LSI does take into account the frequency of words in the documents, whereas WE with negative sampling does not

Question **12**

Not yet answered

Marked out of 1.00

When computing *PageRank* iteratively, the computation ends when...

Select one:

- a. The norm of the difference of rank vectors of two subsequent iterations falls below a predefined threshold
- b. The difference among the eigenvalues of two subsequent iterations falls below a predefined threshold
- c. The probability of visiting an unseen node falls below a predefined threshold
- d. All nodes of the graph have been visited at least once

◀ Midterm

Jump to...